Data Analysis, Imaging, and Visualization (DAIV)

Chris Johnson
Nagiza Samatova

August 17-19, 2009

Panel Members

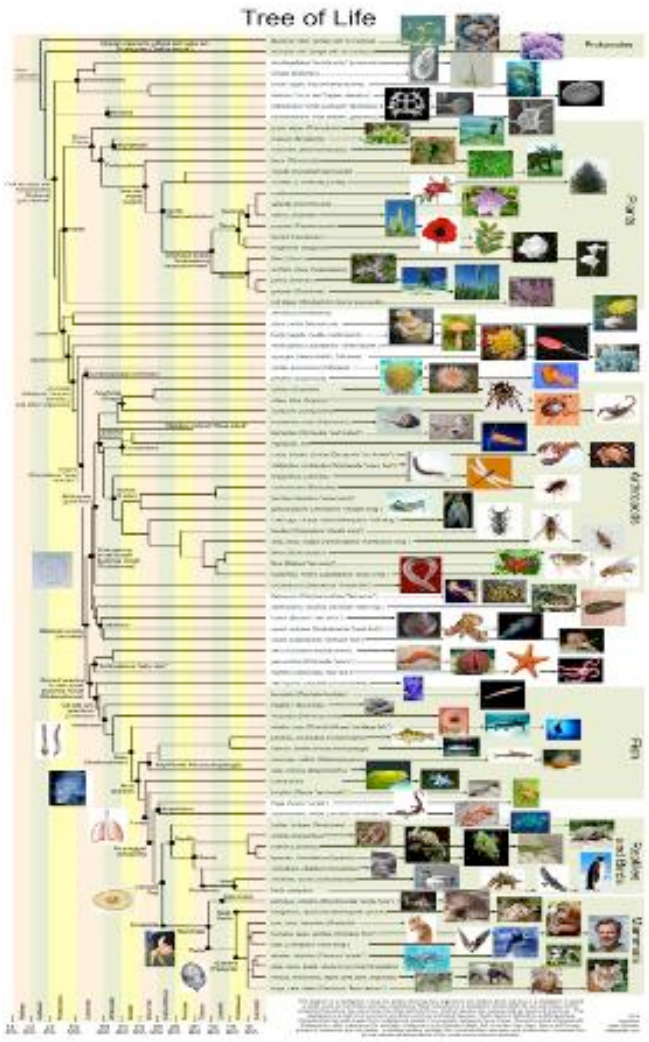
- Gabrielle Allen, gallen@cct.lsu.edu (LSU)
- Bill Cannon, william.cannon@pnl.gov (PNNL)
- Ian Foster, foster@anl.gov (ANL)
- Garth Gibson, garth@cs.cmu.edu (GMU)
- Chris Johnson, crj@sci.utah.edu (U. Utah)
- Tony Hey, tonyhey@microsoft.com (Microsoft)
- Albert Lawrence, albert.rick.lawrence@gmail.com (UCSD)
- Mark LeCros, malecros@lbl.gov (LBNL)
- Michael Papka, papka@anl.gov (ANL)
- Nagiza Samatova, samatovan@ornl.gov (NCSU+ ORNL)
- Tolda Tasdizen, tolga@sci.utah.edu (U. Utah)
- Michael Thelen, mthelen@llnl.gov (LLNL)
- Lora Wolfe, lwolf@anl.gov (ANL)

Top Cross Cutting Computational Problems

- Image Analysis
- Visualization
- Data Analysis
- Data Management
- Workflows
- Social, Economic, Political, & Educational Issues

Phylogenetic Tree of Life for 10-100 Million Species

Computing the optimal phylogenetic tree based on the entire genome of 10 species is intractable even with peta-scale systems



Complexity

Maximum likelihood

C~=$n^{2*}m$ where

n is number of contemporary species

m is size of the genome in question

Requirements

For n=10^6 species and

m~1000 pairs or n=1000

species on the entire genome:

Runtime: 1-3h at 1PetaFLOP

Memory: ~3 TB

Ab initio Prediction of Protein 3-D Structures

Each step is an NP-hard combinatorial optimization problem with different search heuristics.

Complexity of Data Exploration Pipelines

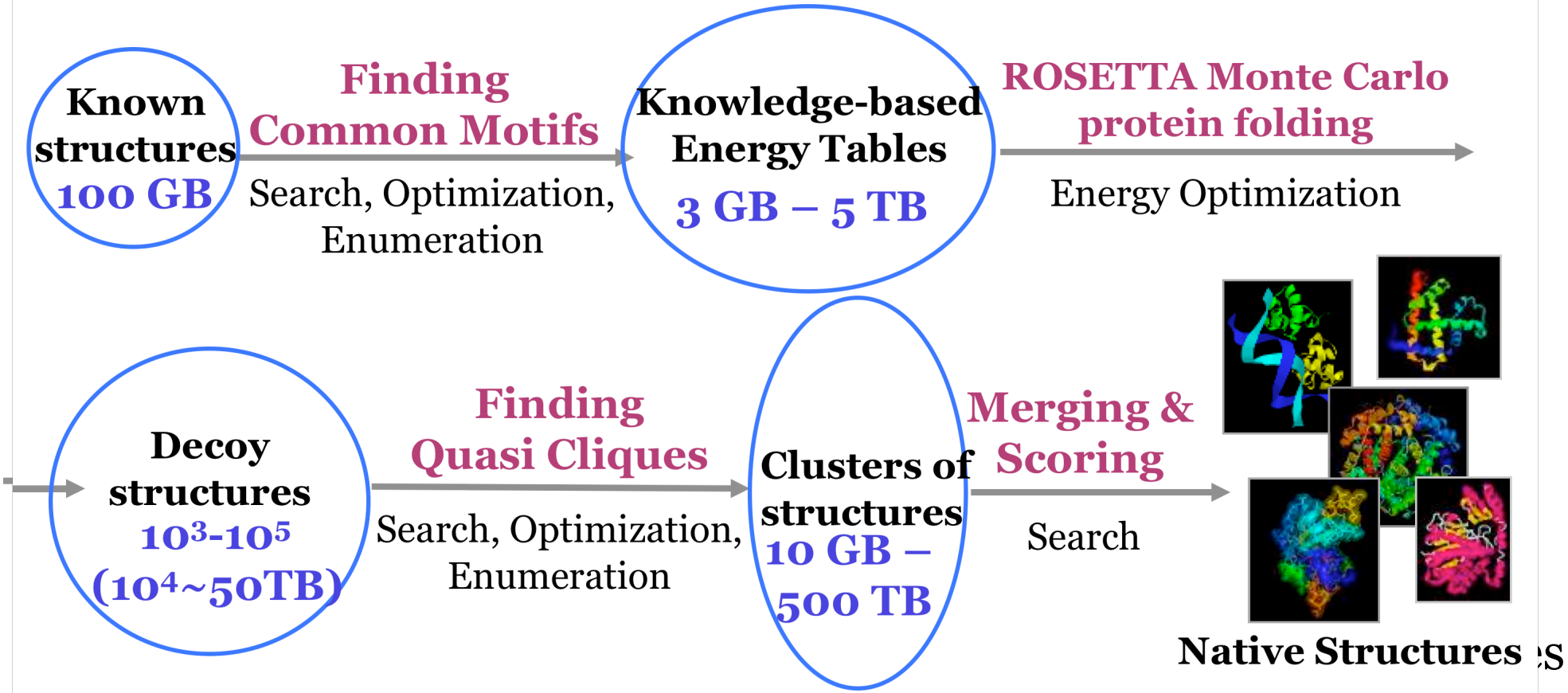


Image Analysis

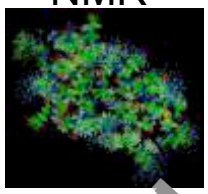
- Relevance to biological drivers
  - Studying phenotypes and populations
  - Tracking cells in time
  - Provide connectivity information for reverse engineering the brain
- Imaging across scales
  - Subcellular to cells to tissues to organs
  - Electron microscopy
  - Confocal & multiphoton microscopy
  - Light microscopy
  - DTI, MRI, fMRI
- Major problems in image analysis that need better solutions
  - Registration of many images, possibly of different modalities at very different scales
  - Automatic segmentation of objects of interest: cells, subcellular structures etc.

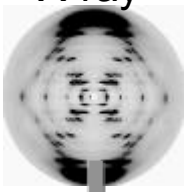
Image Analysis

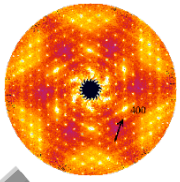
- Challenges
  - Example: Imaging data for entire mouse brain at EM resolution will constitute 30 Peta Voxels
  - Data management
  - Image registration
    - Assembling extremely large volumes
    - Multimodal registration across different modalities
  - Visualization and human annotation
    - Google maps or 3D visualization?
    - New display technologies
  - Robust automatic segmentation and identification of cells and sub-cellular structures
    - Human in the loop to teach the computer: supervised learning
    - Need petaflops to take advantage of all available human annotation
    - Image analysis is largely local: Parallel algorithms for training classifiers on peta-voxel size image data

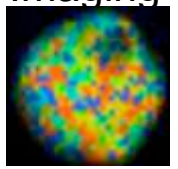
Data-driven Predictive Model Building


Genomics

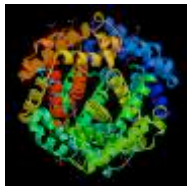

NMR

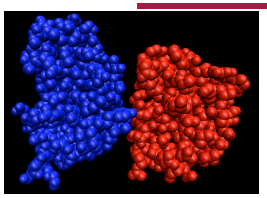

X-ray

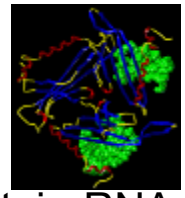

Neutron Scattering

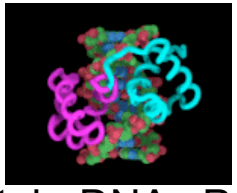

Imaging

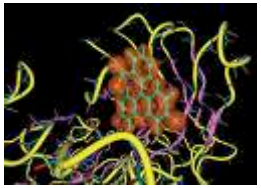
Structural Models


3-d Structure

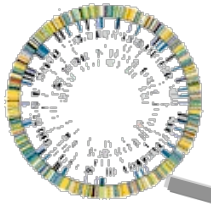

Protein Docking


Protein-RNA

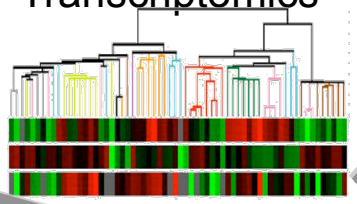

Protein-DNA


Protein-Ligand

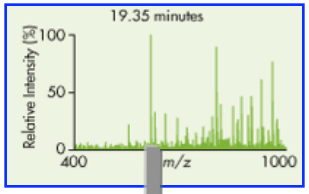
Genomics



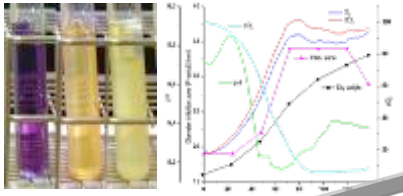
Transcriptomics



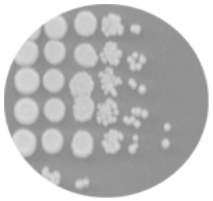
Quantitative Proteomics



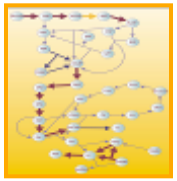
Metabolomics



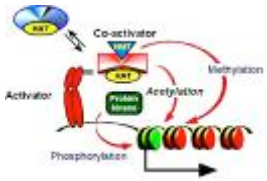
Interactomics



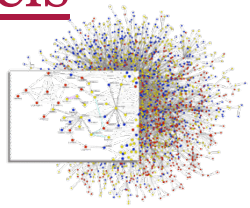
Network/Pathway Models


Metabolic


Regulatory

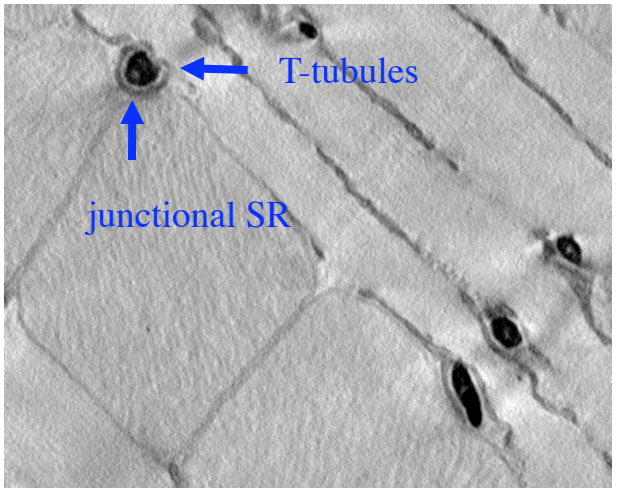


Signaling

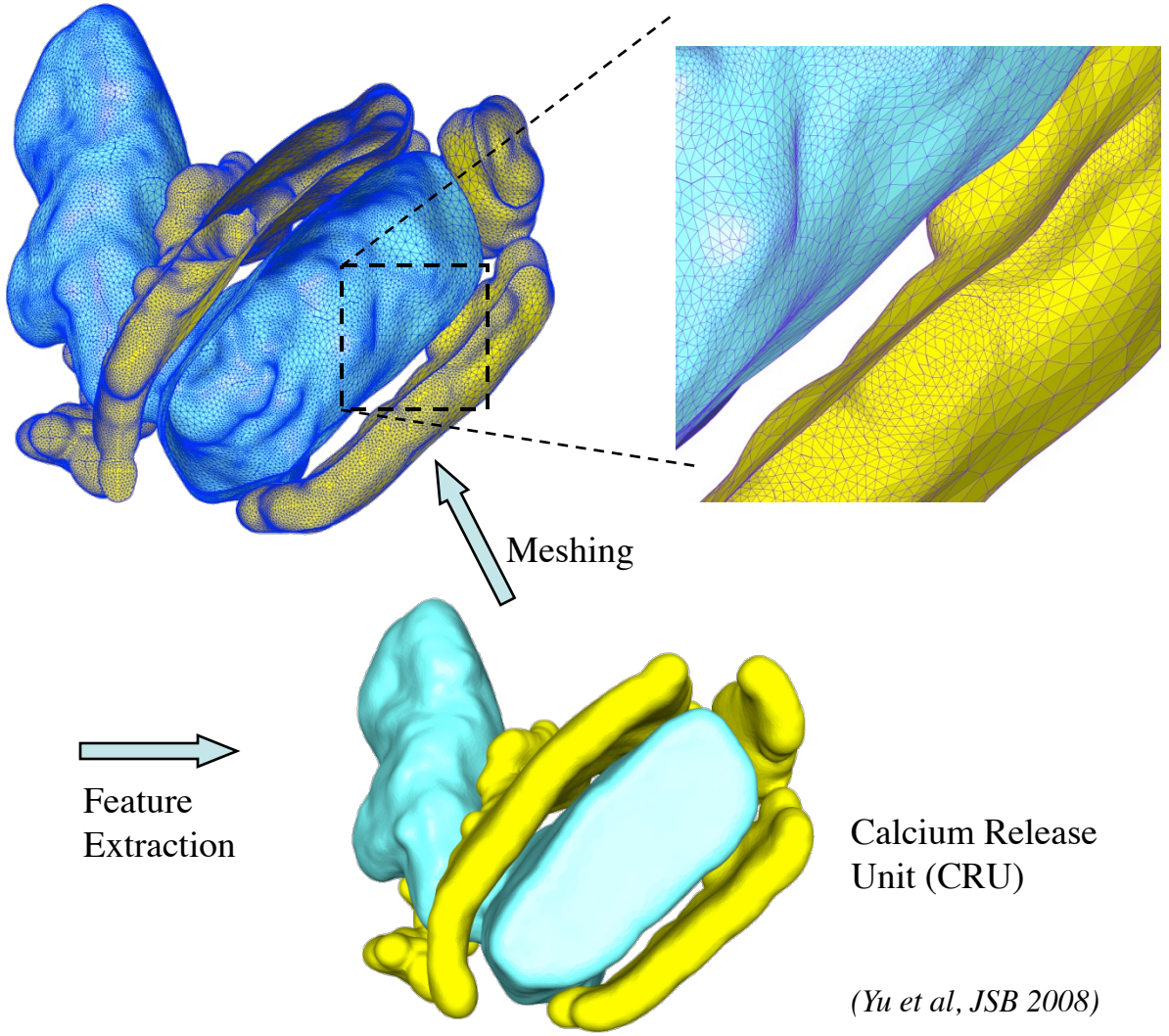

Protein Interaction

GAMer: Building Geometric Models from Electron Microscope Data

JEM-4000EX IVEM (400 kV) (NCMIR, UCSD)



100 nm Image courtesy: Masahiko Hoshijima (UCSD)



Personnel: Zeyun Yu (lead), Michael Holst.
Expected Outcomes: Improved algorithms/software for mesh generation

http://www.FETK.org

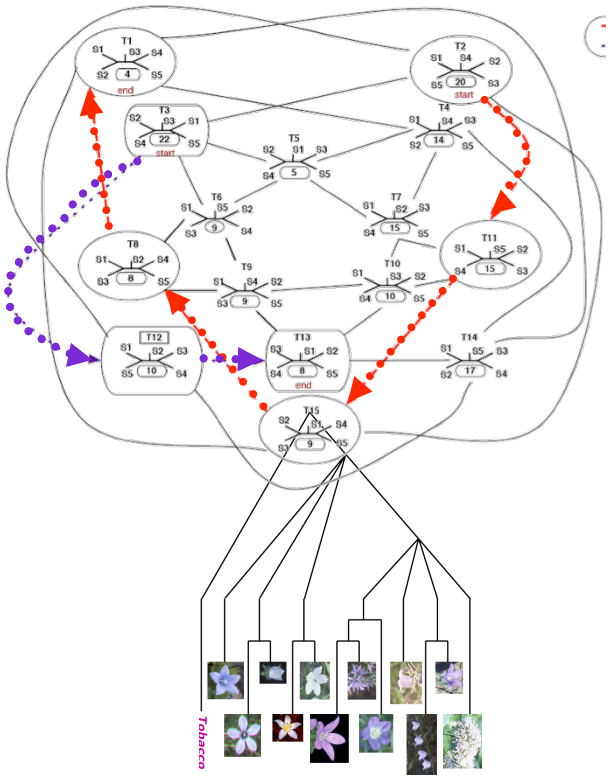
Visualization

- Strong connection to underlying technologies
  - Data management
  - Workflows
- Data fusion
  - Integration of multiple data sources at different scales in both space and time
- Information Visualization
  - Abstract data layout
  - Representation
  - Interaction
- Scientific Visualization
  - In situ
  - Interaction

Challenges: Visual Exploration of the Search History

Exascale Questions:

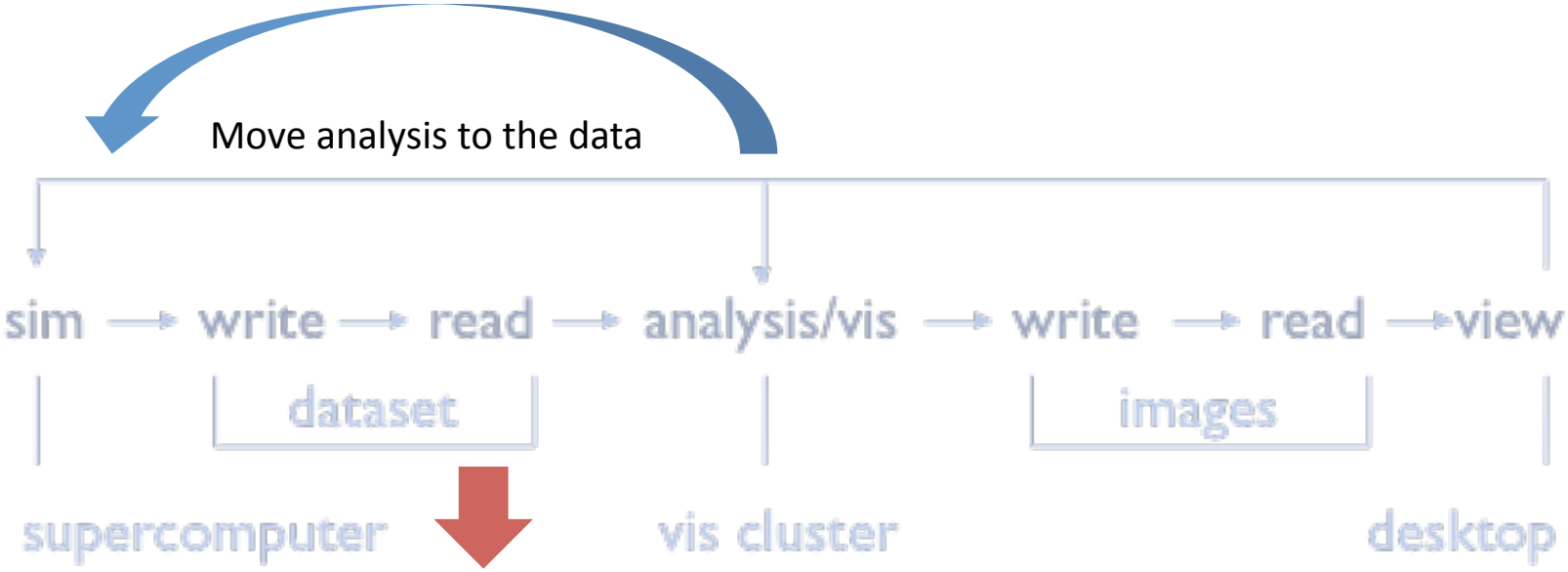
- How to explore the landscape of local optimalities (search histories)?
- How to compare the search histories from different heuristics?
- How to align two/many trees?
- How to visualize the hierarchical clusters of trees?



Impact

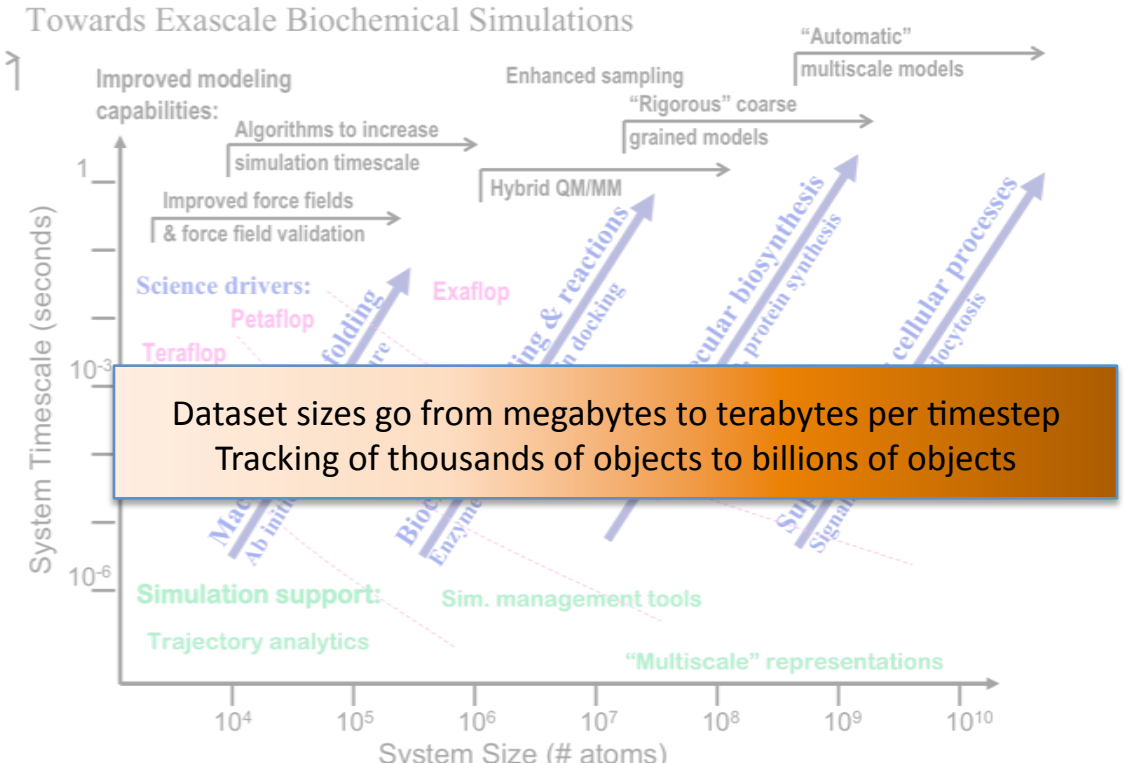
Design of better heuristics
More accurate reconstruction of phylogenies

In situ Visualization



Write smaller, selected results (knowledge bytes)

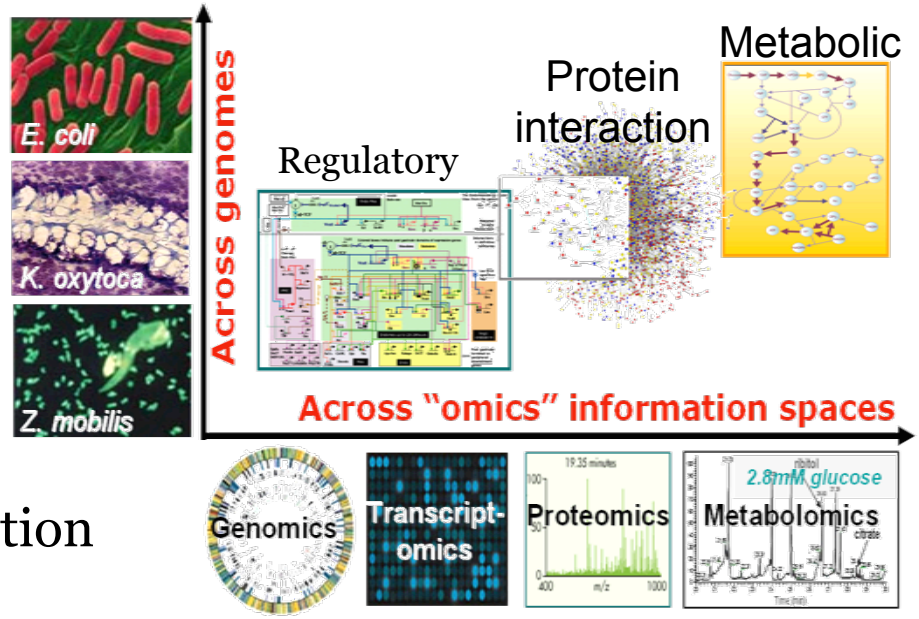
- Integrate analysis and visualization with running simulation
- Exploit additional information available at runtime
- Reduce I/O footprint



Visual Exploration of Networks Evolution

Ultrascale Questions:

- What network motifs are evolutionary conserved?
- Is the conservation statistically significant (compared to random networks)?
- Is a network motif of interest evolutionary conserved? Across what organisms? Are these organisms evolutionary close or distant?
- How to visually compare networks across organisms and “omics” information spaces?

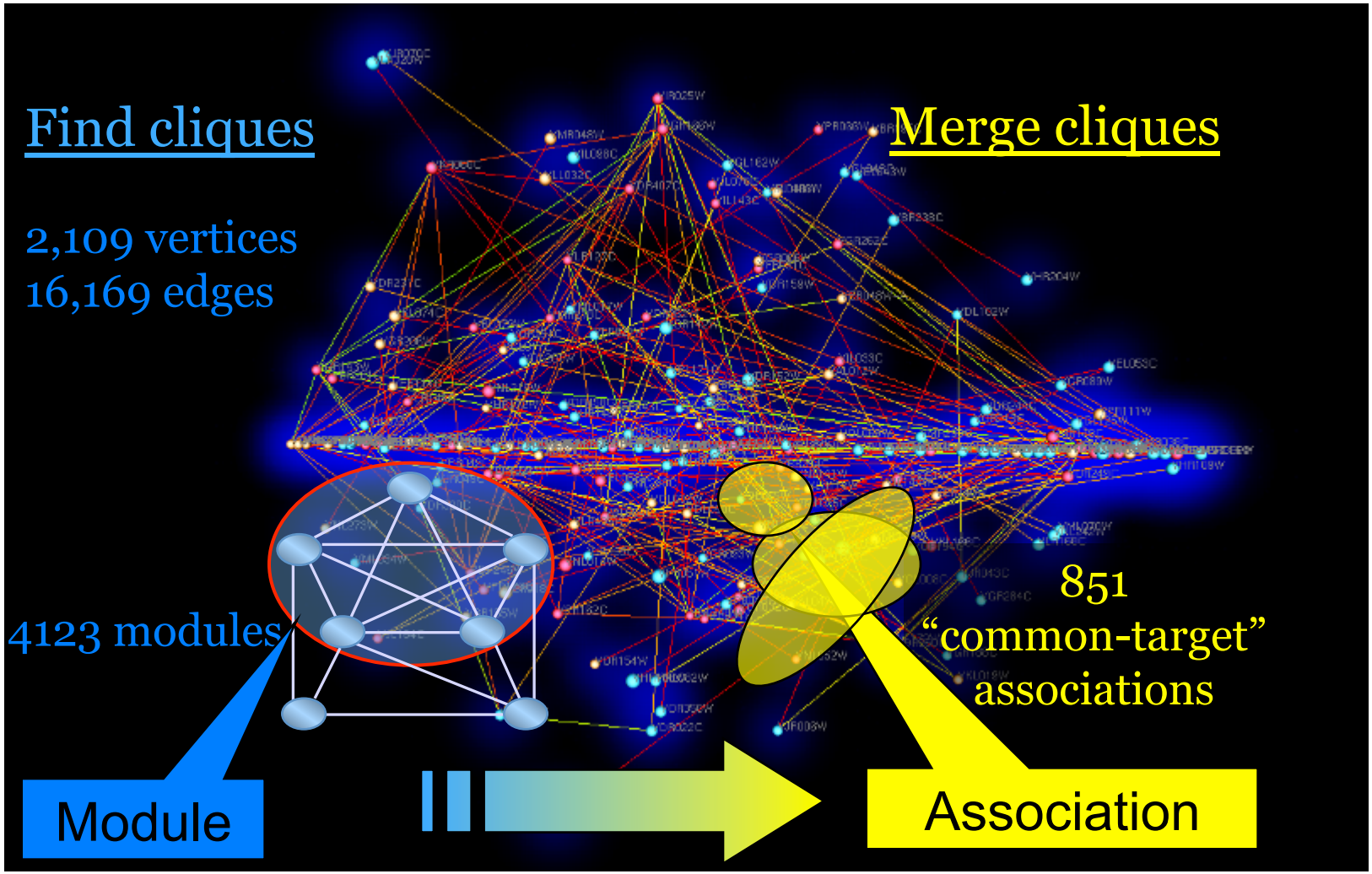


Impact

Design better network analysis
Discover novel network motifs
Annotate proteins with unknown function

Visual Exploration of Genome-scale BioNetworks

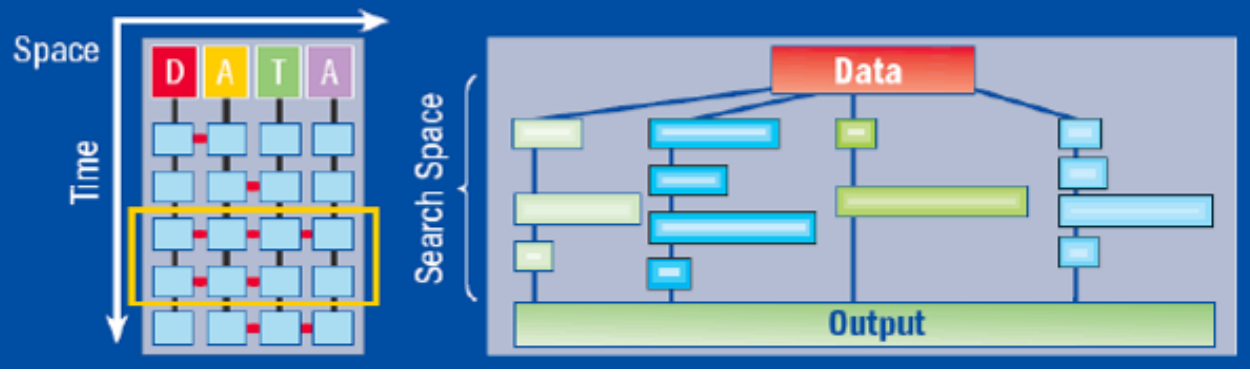
Visual exploration of bionetworks requires solving NP-hard graph problems(e.g., clique/quasi-clique enumeration).



Distinct Data Access Patterns

Data-driven model building presents data-intensive search and enumeration challenge and requires a different mix of memory, disk storage, & communication trade-offs.

	Simulation	Search	Enumeration
Input	Medium	Huge	Medium
Memory Access	Local (2 Time Steps)	Global (Entire Database)	Exponential Irregular
Output to Disk	Iterative	Irregular	Irregular, Huge
Communication	Intensive	For Scoring	Load Balancing
Arithmetic	Float	Integer (Float)	Integer (Float)



Exascale Information Integration and Mining

- Faster-than-Moore's-Law growth in data volumes is leading to an increasing focus on multi-modal data mining and integration as a means of discovery in biology. Thus, we see potential for using exascale computers as massive information integrating and mining tools. Integrating across all extant biological literature, what can we learn about what is known, how knowledge has evolved over time, where inconsistencies appear to exist? By integrating across all extant biological data, can we build maps of the known that make clear where different data agree, where they disagree, where they support published conclusions and where do they not? Can we then use these maps of the literature and data to help define further experiments and/or formulate computational models that may help reduce inconsistencies and broaden the scope of what is known? These questions are scientifically challenging. They also imply a need for much progress on algorithms, programming models, systems software, and potentially also hardware systems.

Workflows, Provenance

- Human-in-the-loop: Interactive steering of complex multi-component models, real time computational exploration (e.g. using exascale computers to solve petascale problems in real time), simple and intuitive interfaces that can aggregate information (e.g. from 10^18 cores) for decision making.
- Annotation: Crucial at all levels as we move to exascale computing and more automated workflows and tools. Annotation of components, data, etc needed that can enable provenance, high level scripting of workflows, validation, data archiving etc.
- Exascale Models: Need to abstract scientific codes from new exascale technologies which will address issues such as Fault tolerance (e.g. checkpointing will not be the paradigm), Parallelism (e.g. message driven paradigms), Architecture heterogeneity (e.g. accelerators, memory hierarchies). This can be achieved through abstraction layers in high level domain languages, component frameworks and scientific libraries. Addressing challenges faced by complexity in the data structures, multiscale, multiphysics, nature of models will be crucial.

Workflows, Provenance

- Power:Exascale computing will be power constrained. Power consumption needs to be taken into account in decision making about the scope and scheduling of workflows. Power-based criteria for optimizations of data analysis workflow.
- Data: File based I/O is not expected to scale from petascale to exascale, visualization and analysis may need to be entirely in -situ. This will totally change the nature of models, run and the analyze will not be possible, models will need to be contain more intelligence to perform the appropriate validation, analysis etc at run time.

Innovative programming models and systems software to support data-intensive and many task applications

- Many--arguably most--problems in biology involve the analysis of large quantities of data and/or many loosely coupled "many-task" computations. These problems require more than a good MPI or MPI/OpenMP implementation and support for parallel I/O: they require extreme-scale data-intensive computing and scripting (aka workflow) paradigms, and runtime, I/O, and operating system methods (and perhaps also hardware systems) that can support many concurrent independent operations and dataflow coordination between activities with efficient support for {one,few,many} x {reader,writer} coordination patterns.

Semantic Technologies for Exascale Problems

- One of the Exascale problems facing biology is intelligent text mining to extract semantic information from the huge and growing literature. Hand annotation of both textual information and experimental data can only hope to reach a tiny percentage of the literature and data currently 'published'. Tools and technologies to automate the extraction of semantic information and to annotate both text and data are a prerequisite for ensuring that we can build on existing experiments. Some of this data will come from large scale simulations and will need to be compared and combined with experimental data in a 'scientific mash-up'. Agreement by the different research communities on ontologies and data formats for exchange and reuse will be important components of any intelligent cyberinfrastructure for biological research. Just as there is a social issue with recognizing computational scientists as a valid discipline worthy of academic rewards, so too, there is a need for recognition of data curators and archivists who make possible the preservation and reuse of data.

Mathematical Challenges

- Mathematical Challenge One: The Mathematics of the Brain Develop a mathematical theory to build a functional model of the brain that is mathematically consistent and predictive rather than merely biologically inspired.
- Mathematical Challenge Two: The Dynamics of Networks Develop the high-dimensional mathematics needed to accurately model and predict behavior in large-scale distributed networks that evolve over time occurring in communication, biology, and the social sciences.
- Mathematical Challenge Three: Capture and Harness Stochasticity in Nature Address Mumford's call for new mathematics for the 21st century. Develop methods that capture persistence in stochastic environments.
- Mathematical Challenge Four: 21st Century Fluids Classical fluid dynamics and the Navier-Stokes Equation were extraordinarily successful in obtaining quantitative understanding of shock waves, turbulence, and solitons, but new methods are needed to tackle complex fluids such as foams, suspensions, gels, and liquid crystals.

Social, Economic, Political, & Educational Issues

- **Hardware and software issues:**
  - Finding: Petascale computers have been mostly utilized for simulations. Very few examples exist for the use of such machines for data analysis and visualization
  - Recommendation: Data analysis, imaging and viz is extreme-scale data- and compute-intensive problem. Need to explore the right hw architectures, programming models and algorithms for such problems
- **Batch, interactive (human-in-the-loop) data exploration:**
  - Finding: At extreme scale, batch processing is a necessity. Yet the algorithms are not mature enough.
  - Recommendation: Investigate and support proper balance between batch mode data exploration and human-in-the-loop/interactive steering and exploration

Social, Economic, Political, & Educational Issues

- **Multi-disciplinary training:**
  - Finding: Extreme scale computing will enable studying the bio systems at much higher level of complexity. This will require a mix of skills in multi-disciplinary science including biology, HPC, statistics, machine learning, scalable algorithms, etc.
  - Recommendation: Initiate a systematic effort for effective training the next generation of investigators for extreme-scale biosciences
- **Single investigator projects vs. big center projects:**
  - Finding: Single investigator projects have difficulty with utilizing the breadth and depth of multi-disciplinary bioscience.
  - Recommendations: Encourage single investigator projects mostly for EARLY CAREER projects. Increase the efficacy of utilizing multi -institutional, multi-disciplinary teams working on grand challenge problems. Ensure sustained funding for productive teams.

#1: Rapid, high fidelity assessment of metabolic, and regulatory potential of 1000s microbes

Scientific and computational challenges/gaps

- Most techniques work with single modality data (e.g. genomics only).
- Most (?) information visualization techniques don't scale to extreme scale of complexity of biological systems.
- Errors in annotations propagate across multiple scales.

Summary of research direction

- Novel algorithms for integrative and comparative analysis and vis. of complex bio data (e.g. network inference)
- Scalable data- and compute-intensive DAIV algorithms
- Uncertainty quantification, community -level-annotation, mapping to scientific literature in a semantically consistent way

Expected Scientific and Computational Outcomes

- Identify inconsistencies and errors in genome-scale annotations
- Facilitate predictive understanding of genotype-phenotype relationships
- Support bioengineering of microbial systems with target phenotypic properties

Potential impact on Biological Science

- Bioenergy: Identify target organisms and their systems properties for enhanced biomass production.
- Bioremediation: Fight corrosion in an environmentally safe manner
- Carbon cycle: Identify key components for efficient photosynthesis

#2: Predict and simulate microbial behavior and response to changing environmental or process-related conditions

Scientific and computational challenges/gaps

- Parameter space is enormous and hard to measure experimentally.
- Increase spatio-temporal resolution, type and use of multi-modal, multi-scale biological imaging: single cell, biofilm,..

Expected Scientific and Computational Outcomes

- Statistically sound designs of model-and simulation-driven experiments
- Create higher resolution, more complete 3D multi-scale bio models/simulations.
- Build parametrized key subcellular metabolic models and their regulation
- Map community structure to biogeochemical function

Summary of research directions

- Inverse problem solvers: inference of many parameters with a few observables
- Advanced analysis of simulation outputs from highly under-determined models
- Statistical methods for experiment design, model validation and verification under huge uncertainty

Potential impact on Biological Science

- Bioenergy: Increase biofuel sustainability through understanding of microbes-plant interaction and nutrients uptake.
- Carbon cycle: Explore biogeochemical response to climate change

#3:Quantitative imaging of macromolecules in single cells in space and time

Scientific and computational challenges

- Integrated analysis of multiple image sources
- Automated large-scale image processing and data management of thousands of cells over space and time

Summary of research direction

- Multi-modal image registration
- Time dependent, semi-automatic parallel segmentation with human in the loop
- High dimensional visualization

Expected Scientific and Computational Outcomes

- Add text here

Potential impact on Biological Science

- Use data to build and test mathematical models of molecular regulatory networks

#4:Reverse engineering of the brain: Neural reconstruction

Scientific and computational challenges

- Algorithmic barriers to large-scale reconstruction of neural circuitry from serial-section TEM: volume assembly and process tracking/synapse detection.
- Create larger models and integrate into multi-scale models

Expected Scientific and Computational Outcomes

- The National Academy of Engineering has selected *reverse engineering the brain as one of their grand challenges with the motivation that part of the problem with state-of-the-art thinking machines is that they have been designed without much attention to real ones*.

Summary of research direction

- Automatic 3D EM image analysis, volume assembly and registration
- Parallel segmentation and annotation with human in the loop
- Create multi-scale models for functional simulation

Potential impact on Biological Science

- Understanding neuro-degenerative diseases and building neural implants such as artificial retinas to cure blindness.
- Design of better, smarter computers

#5:Image-based Phenotyping

Scientific and computational challenges

- Large-scale image analysis
- Creating complex geometric models
- Anatomical and functional analysis between populations of knock outs and wild types

Summary of research direction

- Better, parallel segmentation tools
- Better large-scale meshing tools
- Shape statistics for populations

Expected Scientific and Computational Outcomes

- Fill in later

Potential impact on Biological Science

- Understand physiological instantiation of geneotype for individuals and populations
- Create personalized genetic-based diagnoses and treatments